CS-GY 6763: Lecture 2
Count-Sketch, Union Bound, Exponential Tail Bounds

NYU Tandon School of Engineering
Prof. Rajesh Jayaram

Note on Mathematical Proofs

It can be hard to know how formal to be. We will try to provide feedback on first problem set for anyone who is either too rigorous or too loose. It's a learning process.

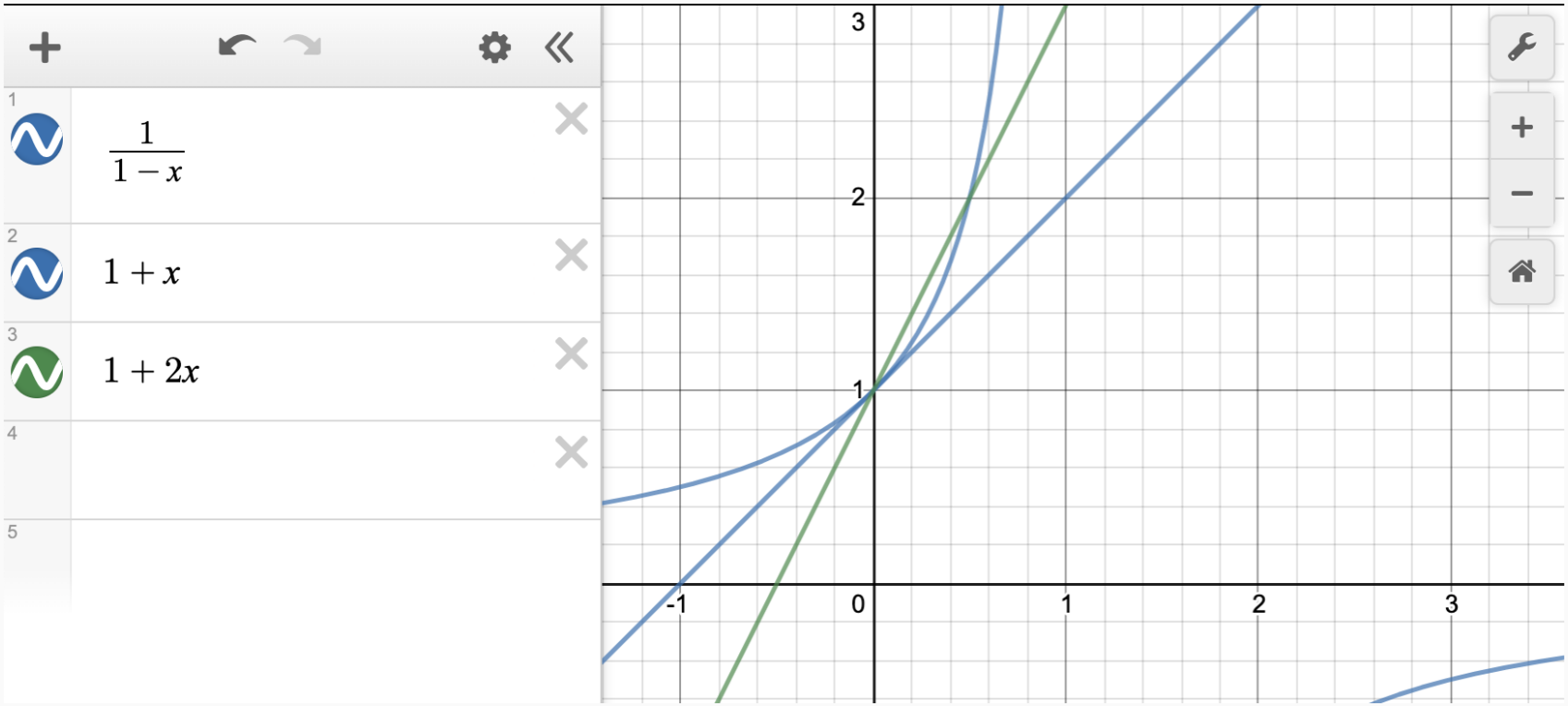
Things that are generally fine:

- Can assume input size n is $> C$ for some constant c. E.g. $n > 2, n > 10$.
- Similarly can assume $\epsilon < c$ for constant c. E.g. $\epsilon < .1$, $\epsilon < .01$.
- If I write $O(z)$, you are free to choose the constant. E.g., it's fine if your method only works for tables of size $1000 \cdot m^{1.5}$.
- Derivatives, integrals, etc. can be taken from e.g. WolframAlpha without working through steps.
- Basic inequalities can be used without proof, as long as you verify numerically. Don't need to include plot on problem set.

Example inequality

$$1 + \epsilon \leq \frac{1}{1-\epsilon} \leq 1 + 2\epsilon \text{ for } \epsilon \in [0, .5].$$

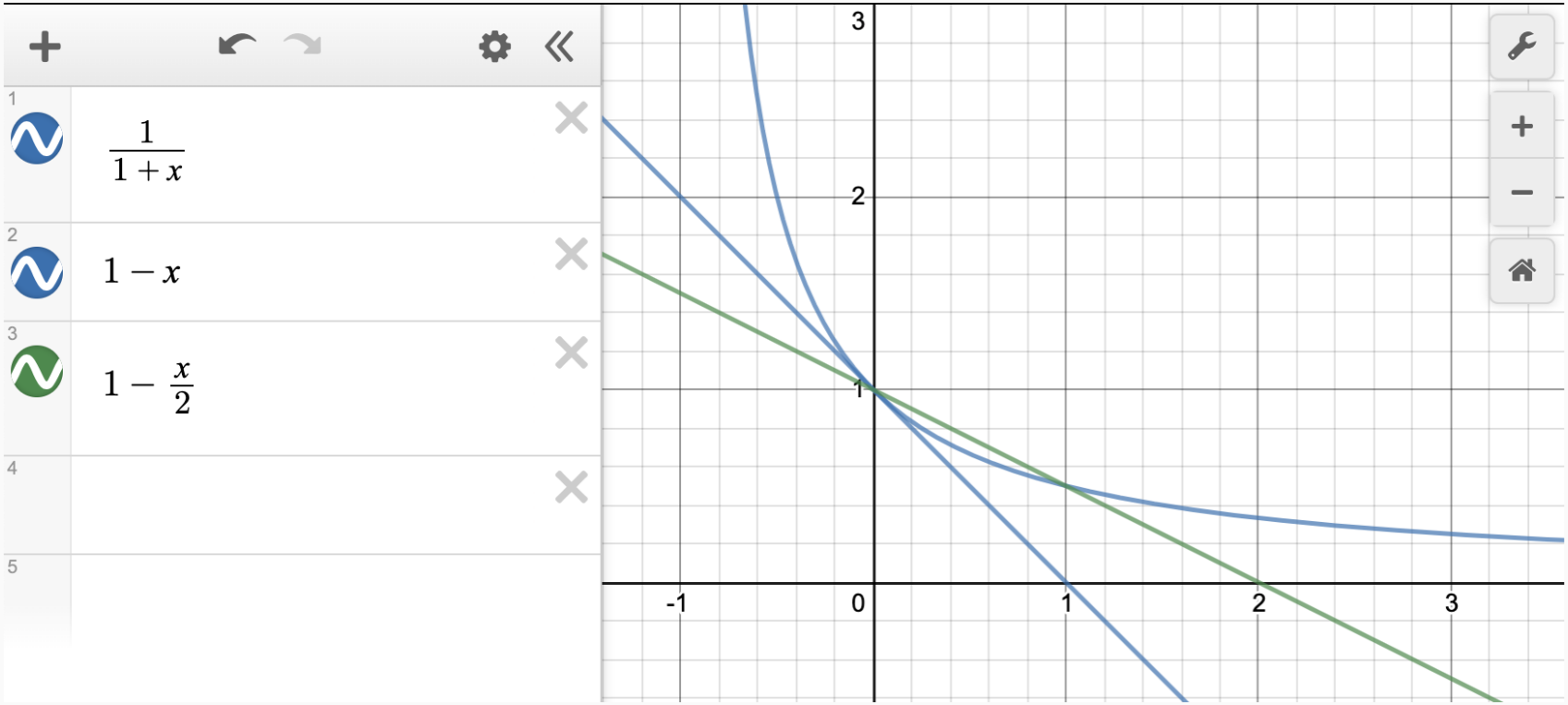
Proof by plotting:



Example inequality

$$1 - \epsilon \leq \frac{1}{1+\epsilon} \leq 1 - .5\epsilon \text{ for } \epsilon \in [0, 1].$$

Proof by plotting:



General Advice

Tip: When confronted with a complex expression, try to simplify by using big-Oh notation, or just rounding things off. Then clean-up your proof after you get to a solution.

Examples:

- $(m-1) \approx m$
- $\frac{1}{n} - \frac{1}{n^2} \approx \frac{1}{n}$
- $\left(\frac{m-1}{cm^{1.5}}\right)^2 \approx O\left(\frac{1}{m}\right)$.
- $\log(n/2) \approx \log(n)$

Link to useful inequalities posted on website.

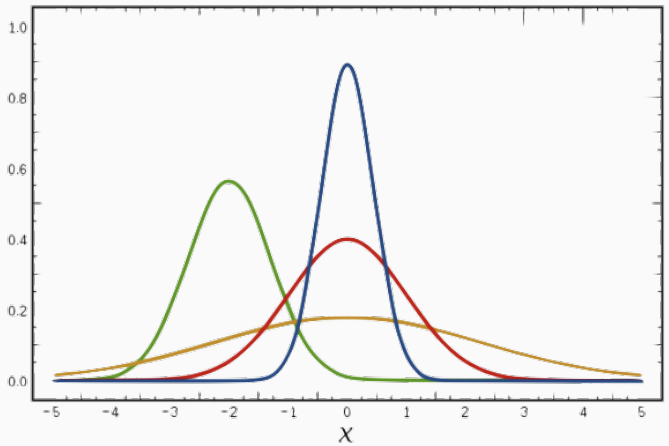
Review of Chebyshev's inequality

A new concentration inequality:

Lemma (Chebyshev's Inequality)

Let X be a random variable with expectation $\mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}[X]$. Then for any $k > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq k \cdot \sigma] \leq \frac{1}{k^2}$$



$\sigma = \sqrt{\text{Var}[X]}$ is the standard deviation of X. Intuitively this bound makes sense: it is tighter when σ is smaller.

Linearity of variance

$E[X_1 X_2] = E[X_1] E[X_2]$ if indep.

Fact: For pairwise independent random variables $X_1, \ldots, X_m$,

$$\text{Var}[X_1 + X_2 + \ldots + X_m] = \text{Var}[X_1] + \text{Var}[X_2] + \ldots + \text{Var}[X_m].$$

I.e., we require that for any i, j X_i and X_j are independent.

This is strictly weaker than mutual independence, which requires that for all possible values $v_1, \ldots, v_k$,

$$\Pr[X_1 = v_1, \ldots, X_k = v_k] = \Pr[X_1 = v_1] \cdot \ldots \cdot \Pr[X_k = v_k].$$

Quick example

If I flip a fair coin 100 times, show that with 93% chance I get between 30 and 70 heads?

Let $C_1, \ldots, C_{100}$ be independent random variables that are 1 with probability $1/2$, 0 otherwise.

Let $H = \sum_{i=1}^{100} C_i$ be the number of heads that get flipped.

$\mathbb{E}[H] =$ 50

$E[(\sum C_i)^2]$

$C_i^2 = \begin{cases} 1^2 \text{ w.p. } \frac{1}{2} \\ 0^2 \text{ otherwise} \end{cases}$

$\text{Var}[H] =$ $\sum \text{Var}(C_i) = \sum_i E[C_i^2] - \sum_i E[C_i]^2$

$50 - 25$

$= 25$

Quick example

If I flip a fair coin 100 **times, show that with** 93% **chance I get between** 30 **and** 70 **heads?**

Let $C_1, \ldots, C_{100}$ be independent random variables that are 1 with probability $1/2$, 0 otherwise.

Let $H = \sum_{i=1}^{100} C_i$ be the number of heads that get flipped.

$\mathrm{Var}[H] = 25$. $= \sigma^2$

Chebyshev's: $\Pr[|H - \mu| > 20] < \frac{1}{16}$

$\mu = 50$

4.5

4.8

Lecture road map

So far, we have seen the power of

- Linearity of Expectation + Markov's Inequality
- Linearity of Variance + Chebyshev's Inequality

Today, we will discuss one of the most powerful tools in all of randomized algorithms:

Union Bound + Exponential Tail Bounds

These six simple tools form the cornerstone of randomized algorithm design.

The Turnstile Streaming Model

Definition (Streaming Model)

Let $f \in R^n$ be the implicit frequency vector, initialized to $\vec{0}$. A *turnstile* data stream is a sequence of updates $(i_1, \Delta_1), (i_2, \Delta_2), \ldots, (i_m, \Delta_m)$, where $i_t \in [n]$ and $\Delta_t \in \mathbb{Z}$. The update (i_t, Δ_t) causes the change

$$f_{i_t} \leftarrow f_{i_t} + \Delta_t$$

Note that updates Δ_t can be negative. This is harder than the insertion only model!

- Differences between streams (i.e. $f :=$ difference between IP traffic sent through router A vs router B).
- Updates to high-dimensional gradients $f = \nabla g$ in SGD and other optimization methods

Heavy Hitters – more formally

$$\|f\|_p = \left(\sum_i |f_i|^p\right)^{1/p} \qquad p > 0$$

Definition (ϵ-Heavy Hitters Problem)

Consider a stream of m updates $(i_1, \Delta_1), (i_2, \Delta_2), \ldots, (i_m, \Delta_m)$, resulting in a frequency vector $f \in \mathbb{R}^n$. Return a set $S \subset [n]$ containing all indices i such that $|f_i| \geq \epsilon \|f\|_1$, and no i such that $|f_i| \leq \frac{\epsilon}{2}\|f\|_1$.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
5	-12	3	3	-4	5	5	10	0	-3

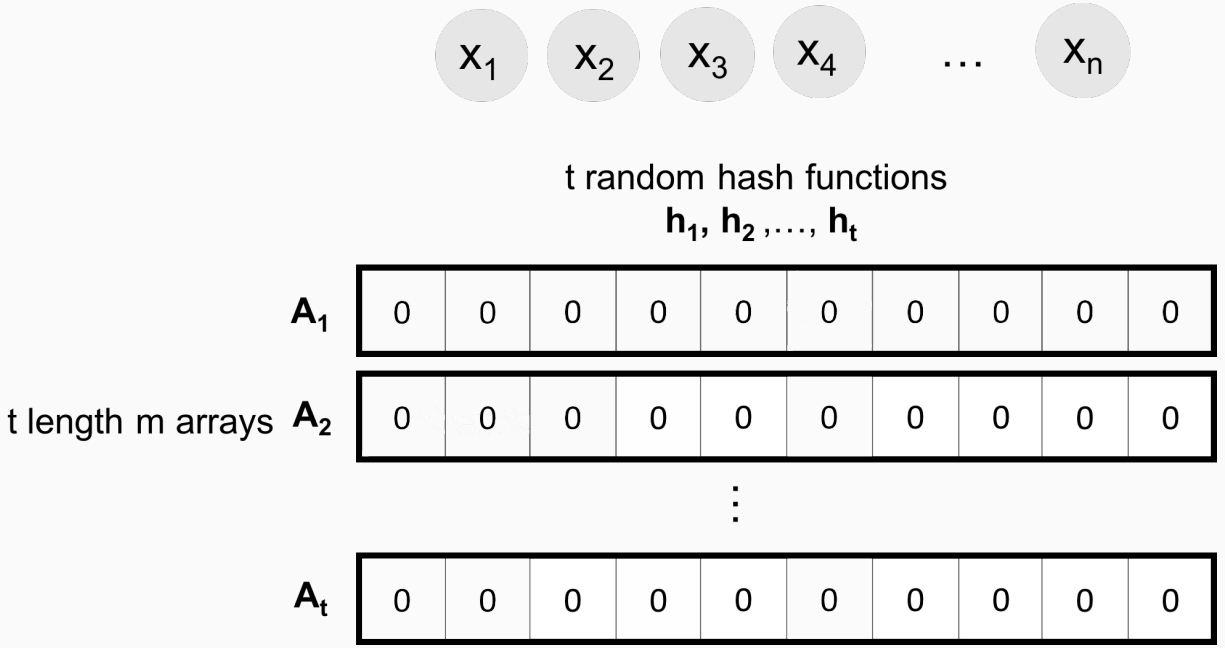
This generalizes the problem from last class, when we were promised that $f_i \geq 0$ for all $i \in [n]$

Recall Count-Min Sketch

Count-Min Update:

- Choose random hash functions $h_1, h_2, \ldots, h_t : [n] \to [B]$
- For each update $\ell = 1, \ldots, m$
  - Given update (i_ℓ, Δ_ℓ), for each $j = 1, 2, \ldots, t$ set

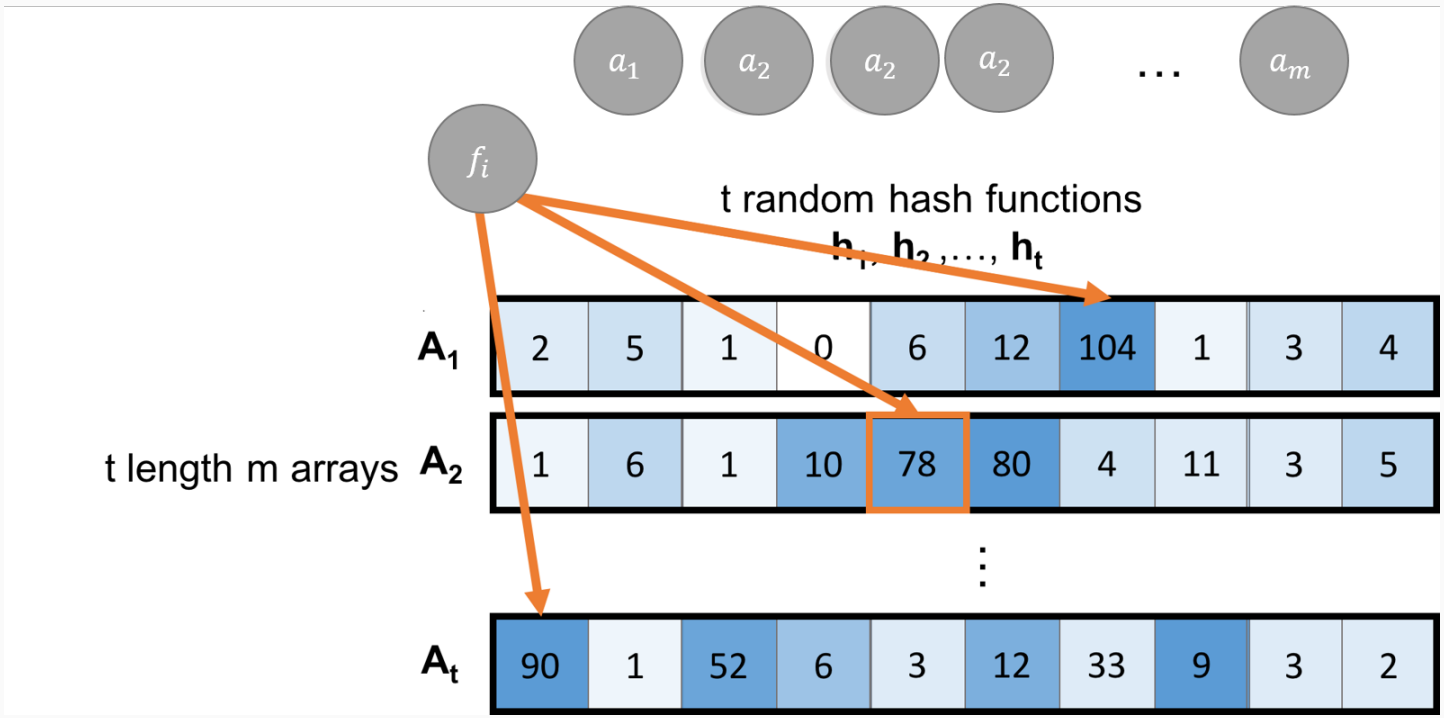
$$\mathbf{A}_j[h_j(i_\ell)] = \mathbf{A}_j[h(i_\ell)] + \Delta_\ell$$



Count-Min Sketch

Estimate of count-min:

$$\tilde{f}_i = \min_{j \in [t]} h_j(i)$$



Count-Min Sketch

Theorem

For any $\epsilon, \delta \in (0,1)$, when run on an insertion-only stream (i.e., $\Delta_\ell \geq 0$ for all $\ell \in [m]$), for any index $i \in [n]$ Count-min sketch yields an estimate $\tilde{f}_i$ of the frequency f_i satisfying:

$$f_i \leq \tilde{f}_i \leq f_i + \epsilon \|f\|_1$$

with probability $\geq 1 - \delta$, using $O\left(\log(1/\delta) \cdot \frac{1}{\epsilon}\right)$ words of space.

- Analysis computed expectation and used Markov's inequality.

Count-Min Sketch Accuracy

Value of a Bucket

$$\mathbf{A}[h(i)] = f_i + \underbrace{\sum_{j \neq i} \mathbb{1}[\mathbf{h}(j) = \mathbf{h}(i)] \cdot f_j}_{\text{error in frequency estimate}}$$

Expected Error:

$$\mathbb{E}\left[\sum_{j \neq i} \mathbb{1}[\mathbf{h}(j) = \mathbf{h}(i)] \cdot f_j\right] \leq \frac{\sum_{i \in [n]} |f_i|}{B} = \frac{\|f\|_1}{B}$$

Count-Min Sketch Accuracy

$$\mathbf{A}[h(i)] = f_i + \underbrace{\sum_{j \neq i} \mathbb{1}[\mathbf{h}(j) = \mathbf{h}(i)] \cdot f_j}_{\text{error in frequency estimate}}$$

Expected Error:

$$\mathbb{E}\left[\sum_{j \neq i} \mathbb{1}[\mathbf{h}(j) = \mathbf{h}(i)] \cdot f_j\right] \leq \frac{\|f\|_1}{B}$$

Markov's inequality: $\Pr\left[\sum_{j \neq i} \mathbb{1}[\mathbf{h}(j) = \mathbf{h}(i)] \cdot f_j \geq \frac{2\|f\|_1}{B}\right] \leq \frac{1}{2}$

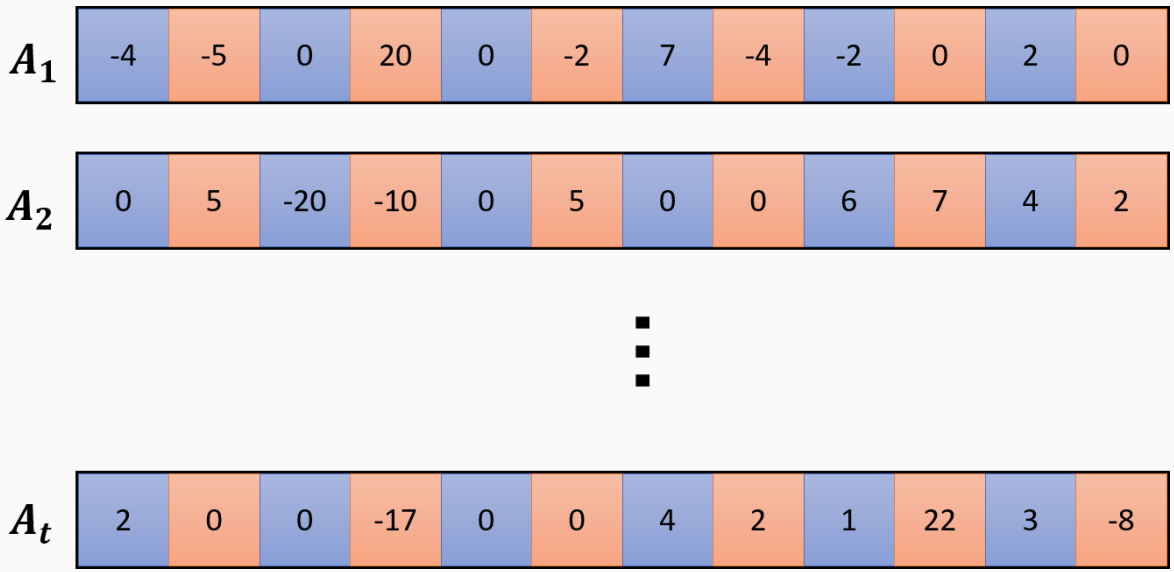
Where does this proof fail for turnstile streams?

Enter the Matrix: Count-Sketch

Count-Sketch Update:

- Choose random hash functions $h_1, h_2, \ldots, h_t : [n] \to [B]$, and $\sigma_1, \sigma_2, \ldots, \sigma_t : [n] \to \{1, -1\}$
- For each update $\ell = 1, \ldots, m$
  - Given update (i_ℓ, Δ_ℓ), for each $j = 1, 2, \ldots, t$ set

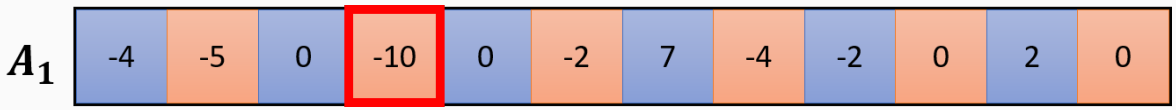
$$\mathbf{A}_j[h_j(i_\ell)] = \mathbf{A}[h(i_\ell)] + \sigma(i_\ell) \cdot \Delta_\ell$$



Enter the Matrix: Count-Sketch

How can we estimate f_i?

$$A_j[h_j(i)] = \sigma_j(i) \cdot f_i + \underbrace{\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k}_{\text{error in frequency estimate}}$$



Expected Error

$$\mathbb{E}\left[\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right] =$$

Enter the Matrix: Count-Sketch

Expected Error

$$\mathbb{E}\left[\sum_{k\neq i} \mathbb{1}[\mathbf{h}(k)=\mathbf{h}(i)]\cdot\sigma_j(k)f_k\right] = \sum_{k\neq i}\frac{f_k}{B}\cdot\mathbb{E}\left[\sigma_j(k)\right]$$

$$= 0$$

How can we show that

$$\left|\sum_{k\neq i} \mathbb{1}[\mathbf{h}(k)=\mathbf{h}(i)]\cdot\sigma_j(k)f_k\right|$$

is not too large?

Back to the Variance

$\mathbb{E}[\text{error}] = 0$. **Variance of Error:**

$$\text{Var}\left[\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right] = \sum_{k \neq i} \text{Var}\left[\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right]$$

$$= \sum_{k \neq i} \mathbb{E}[(\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)])^2] \cdot \mathbb{E}[\sigma_j^2(k) f_k^2]$$

Back to the Variance

$\mathbb{E}[\text{error}] = 0$. **Variance of Error:**

$$\text{Var}\left[\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right] = \sum_{k \neq i} \text{Var}\left[\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right]$$

$$= \sum_{k \neq i} \mathbb{E}[(\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)])^2] \cdot \mathbb{E}[\cdot \sigma_j^2(k) f_k^2]$$

$$= \sum_{k \neq i} \frac{1}{B} f_k^2 \cdot \mathbb{E}[\sigma_j^2(k)]$$

Back to the Variance

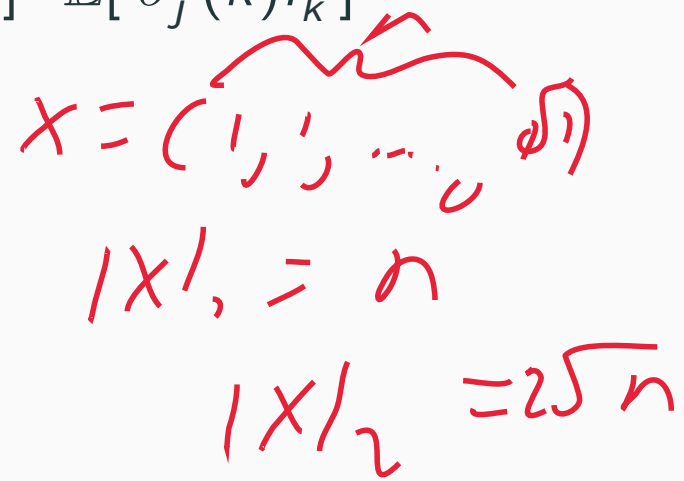
$\mathbb{E}[\text{error}] = 0$. **Variance of Error:**

$$\text{Var}\left[\sum_{k\neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right] = \sum_{k\neq i} \text{Var}\left[\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right]$$

$$= \sum_{k\neq i} \mathbb{E}[(\mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)])^2] \cdot \mathbb{E}[\cdot \sigma_j^2(k) f_k^2]$$

$$= \sum_{k\neq i} \frac{1}{B} f_k^2 \cdot \mathbb{E}[\sigma_j^2(k)]$$

$$= \sum_{k\neq i} \frac{1}{B} f_k^2 = \frac{1}{B}\|f\|_2^2$$



The Variance depends on the L_2 norm of f!

Back to Chebyshev's

$\mathbb{E}[\text{error}] = 0$. **Variance of Error:**

$$\underbrace{\text{Var}[\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k]}_{\text{error in frequency estimate}} \leq \frac{1}{B} \|f\|_2^2$$

Lemma (Chebyshev's Inequality)

Let X be a random variable with expectation $\mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}[X]$. Then for any $k > 0$, $\Pr[|X - \mathbb{E}[X]| \geq k \cdot \sigma] \leq \frac{1}{k^2}$

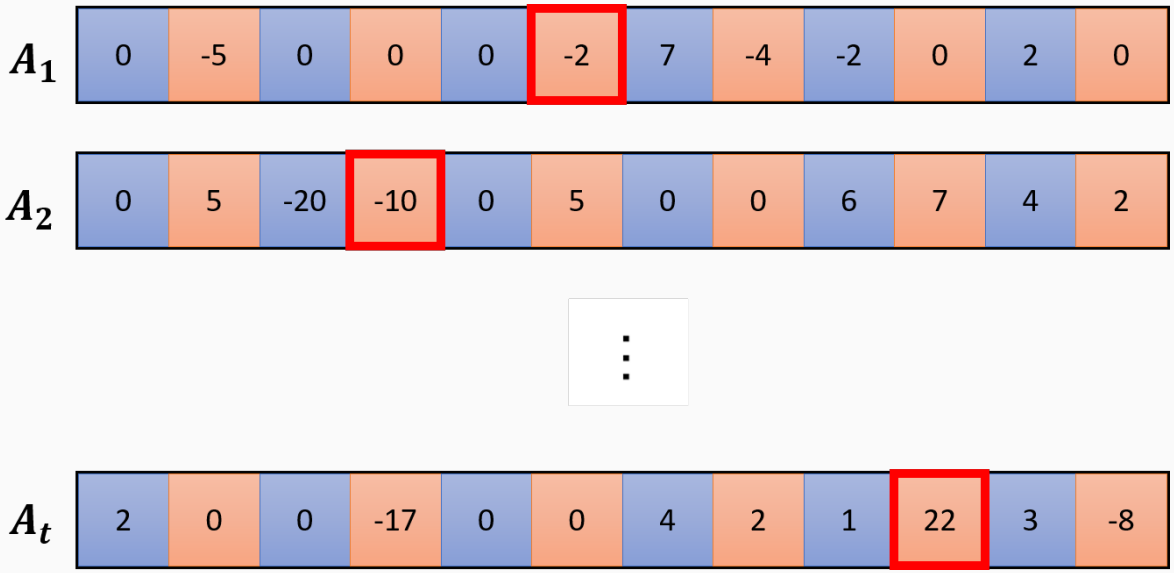
Using Chebyshev's Inequality:

$$\Pr\left[\left|\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k\right| \geq 2 \cdot \frac{1}{\sqrt{B}} \|f\|_2\right] \leq \frac{1}{4} \qquad (1)$$

Enter the Matrix: Count-Sketch

Can we still take the min?

$$\tilde{f}_i = \min_{j \in [t]} A_j[h_j(i)]$$



Variance reduction

$$\mathrm{Var}(cX) = c^2 \mathrm{Var}[$$

Taking min of multiple trials does not work, since error can be negative!

Trick of the trade: Repeat many independent trials and take the mean to get a better estimator.

Given i.i.d. (independent, identically distributed) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2, what is:

- $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$
- $\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \sum \mathrm{var}(X_i)$

$$= \sum E\left[\tfrac{1}{n} X_i^2\right] - E\left[\tfrac{1}{n} X_i\right]^2$$

$$= \frac{1}{n^2} \sum E[X_i^2] - E[X_i]^2$$

$$\frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n} \cdot \sigma^2$$

Variance reduction

Trick of the trade: Repeat many independent trials and take the mean to get a better estimator.

Lemma

Given i.i.d. (independent, identically distributed) random variables $X_1, \ldots, X_n$ with mean μ and variance σ^2, for any $\alpha > 1$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \frac{\alpha}{\sqrt{n}}\sigma\right] \leq \frac{1}{\alpha^2}$$

$\sigma = \frac{\tau \cdot 1 \leq 1/2}{\sqrt{t}} \cdot \sqrt{t} \leq \frac{1}{t} = \delta$

Variance reduction

First Attempt: Repeat many independent trials and take the mean to get a better estimator.

$$\sigma_j(i) \cdot A_j[h_j(i)] = f_i + \sigma_j(i) \underbrace{\sum_{k \neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot \sigma_j(k) f_k}_{\text{error in frequency estimate}}$$

So we can set: $\tilde{f}_i = \frac{1}{t} \sum_{j \in [t]} \sigma_j(i) A_j[\sigma_j(i) h_j(i)]$. Recall that $\text{Var(Error)} \leq \frac{1}{B} \|f\|_2^2$. Setting $B = \Theta(\frac{1}{\epsilon^2})$ and $t = \Theta(\frac{1}{\delta})$ yields:

Lemma

For any **fixed** $i \in [n]$, *the "Count-Mean" Sketch outputs the estimate* $\tilde{f}_i$ *such that*

$$\Pr\left[\left|\tilde{f}_i - f_i\right| \geq \epsilon \|f\|_2\right] \leq \delta$$

Using space $O(\frac{1}{\epsilon^2 \delta})$.

Note on failure probability

$O\left(\frac{1}{\epsilon^2\delta}\right)$ space is an impressive bound, gives good estimates for a single coordinate.

- Achieves any accuracy desired. $1/\epsilon^2$ dependence cannot be improved.
- But... $1/\delta$ dependence is not ideal. For 95% success rate, pay a $\frac{1}{5\%} = 20$ factor overhead in space.
- And this is just to be correct on one cordinate. What if we want output a good estimate for all n coordinates?
- Note that with count-min, we did much better with a $O(\log(1/\delta))$ dependency

We can get a better bound depending on $O(\log(1/\delta))$ using exponential tail bounds.

Why Failure Probabiltiy Matters

Suppose we want to find all coordinates $|f_i| \geq 4\epsilon\|f\|_1$ and no coordinates $|f_i| \leq 2\epsilon\|f\|_1$ (i.e. solve the 4ϵ-heavy hitters problem).

Lemma

For any **fixed** $i \in [n]$, *the "Count-Mean" Sketch outputs the estimate* $\tilde{f}_i$ *such that*

$$\Pr\left[\left|\tilde{f}_i - f_i\right| \geq \epsilon\|f\|_2\right] \leq \delta$$

Using space $O(\frac{1}{\epsilon^2\delta})$.

Since $\|f\|_2 \leq \|f\|_1$, our error $|\tilde{f}_i - f_i| \leq \epsilon\|f\|_2$ is good enough for one coordinate, but we need to be correct on *all* coordinates to solve the heavy hitters problem.

If we are correct for a single $i \in [n]$ with probability $1 - \delta$, what is the probability we are simultaneously correct for all $i \in [n]$?

Bounding a union of events

Goal: Let A_i be the event that our estimate for f_i is bad. In other words

$$A_i := \text{ Event that } |\tilde{f}_i - f_i| > \epsilon \|f\|_2$$

We want to show that none of the A_i's happen. In other words, we want to show:

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] \leq \frac{1}{10}.$$

Need to bound the probability of a union of different events.

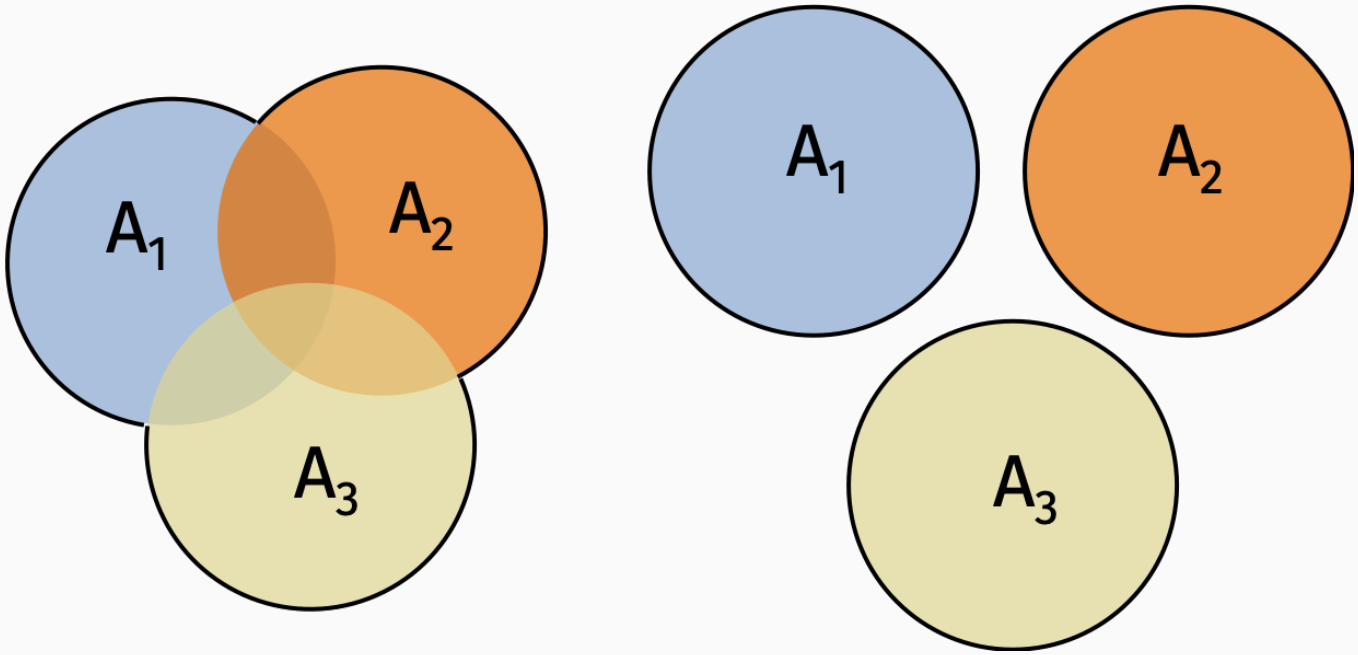
These events are not independent!!

Actually the most important tool in probability

Lemma (Union Bound)

For any random events $A_1, \ldots, A_k$:

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_k] \leq \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_k].$$



Proof by picture.

The Count-Mean Sketch

Let $A_i :=$ Event that $|\tilde{f}_i - f_i| > \epsilon \|f\|_2$. We have

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i] \leq n\delta$$

Lemma

The "Count-Mean" Sketch uses space $O(\frac{1}{\epsilon^2 \delta})$ and outputs $\tilde{f}_i$ such that **for all** $i \in [n]$ *we have:*

$$\left|\tilde{f}_i - f_i\right| \leq \epsilon \|f\|_2$$

with probability at least $1 - \delta n$.

Need to set $\delta < 1/(2n)$ to achieve $> 1/2$ success probability.
This results in $\Omega(n)$ space – which is useless!

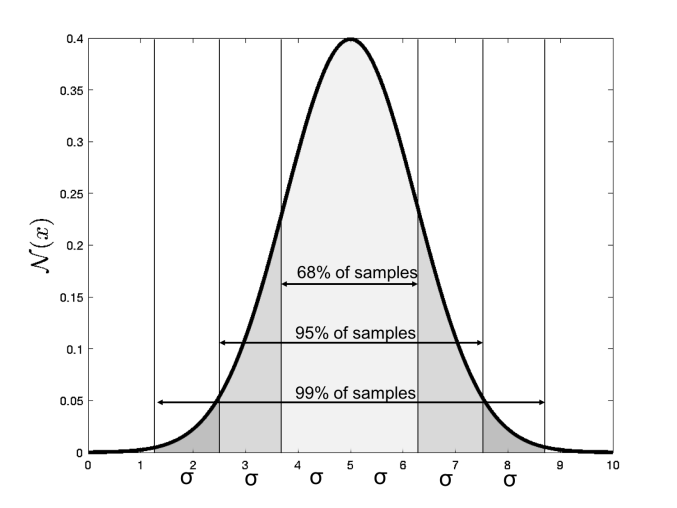
Better Concentration

After the break: Chernoff bounds + Exponential concentration to achieve $\log(1/\delta)$ dependency!

Break

Beyond Chebyshev

Motivating question: Is Chebyshev's Inequality tight?



68-95-99 rule for Gaussian bell-curve. $\mathbf{X \sim N(0, \sigma^2)}$

Chebyshev's Inequality:

$$\Pr(|X - \mathbb{E}[X]| \geq 1\sigma) \leq 100\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 2\sigma) \leq 25\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 3\sigma) \leq 11\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 4\sigma) \leq 6\%.$$

Truth:

$$\Pr(|X - \mathbb{E}[X]| \geq 1\sigma) \approx 32\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 2\sigma) \approx 5\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 3\sigma) \approx 1\%$$
$$\Pr(|X - \mathbb{E}[X]| \geq 4\sigma) \approx .01\%$$

Gaussian concentration

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

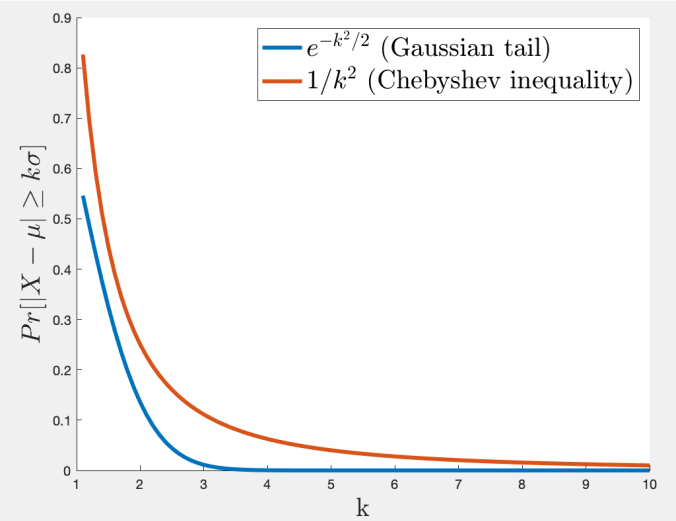
$$\Pr[X = \mu \pm x] = \frac{1}{\sigma\sqrt{2\pi}} \mathbf{e}^{-\mathbf{x}^2/2\sigma^2}$$

Lemma (Guassian Tail Bound)

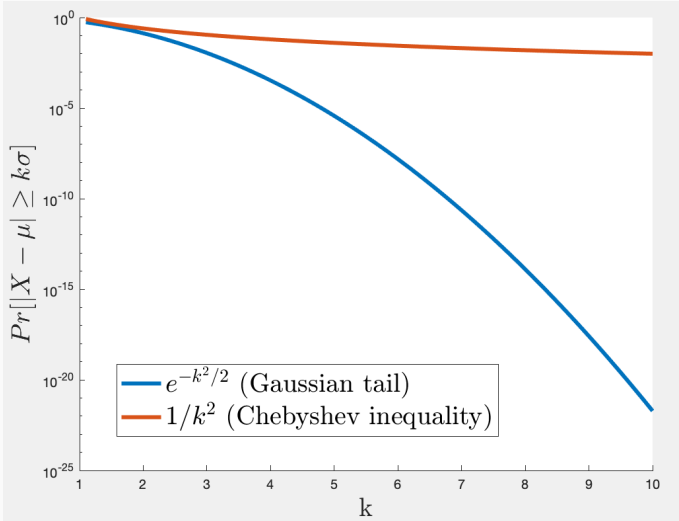
For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\Pr[|X - \mathbb{E}X| \geq k \cdot \sigma] \leq 2e^{-k^2/2}.$$

vs $\frac{1}{k^2}$



Standard y-scale.



Logarithmic y-scale.

Gaussian concentration

Takeaway: Gaussian random variables concentrate much tighter around their expectation than variance alone predicts.

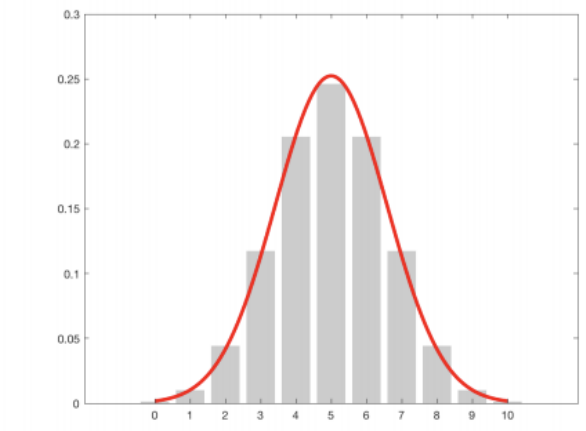
Why does this matter for algorithm design?

Central Limit Theorem

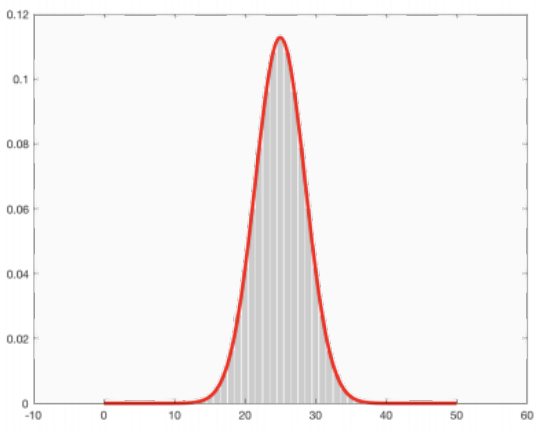
Theorem (CLT – Informal)

Any sum of mutually independent, (identically distributed) r.v.'s $X_1, \ldots, X_k$ with mean μ and finite variance σ^2 converges to a Gaussian r.v. with mean $k \cdot \mu$ and variance $k \cdot \sigma^2$, as $k \to \infty$.

$$S = \sum_{i=1}^{n} X_i \Longrightarrow \mathcal{N}(k \cdot \mu, k \cdot \sigma^2).$$



(a) Distribution of # of heads after 10 coin flips, compared to a Gaussian.



(b) Distribution of # of heads after 50 coin flips, compared to a Gaussian.

Independence

Recall:

Definition (Mutual Independence)

Random variables $X_1, \ldots, X_k$ are mutually independent if, for all possible values $v_1, \ldots, v_k$,

$$\Pr[X_1 = v_1, \ldots, X_k = v_k] = \Pr[X_1 = v_1] \cdot \ldots \cdot \Pr[X_k = v_k]$$

Strictly stronger than pairwise independence.

Exercise

Chebyshev $\Rightarrow \frac{1}{16}$

If I flip a fair coin 100 times, lower bound the chance I get between 30 and 70 heads?

For this problem, we will assume the CLT holds exactly for a sum of independent random variables – i.e., that this sum looks exactly like a Gaussian random variable.

Lemma (Guassian Tail Bound)

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\Pr[|X - \mathbb{E}X| \geq k \cdot \sigma] \leq 2e^{-k^2/2}.$$

4.5

$< 2e^{-8}$

Quantitative versions of the CLT

Lots of different "versions" of exponential concentration bounds

- Chernoff bound
- Bernstein bound
- Hoeffding bound
- Azumas Inequality, McDiarmid's Inequality, Freedman's inequality, Khintshine's inequality, Matrix Chernoff, Matrix Bernstein, Matrix Azuma's ...

Different assumptions on random varibles (e.g. binary vs. bounded), different forms (additive vs. multiplicative error), etc.

Wikipedia is your friend.

Quantitative versions of the CLT

Theorem (Chernoff Bound)

Let $X_1, X_2, \ldots, X_k$ be independent $\{0,1\}$-valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Then the sum $S = \sum_{i=1}^{k} X_i$, which has mean $\mu = \sum_{i=1}^{k} p_i$, satisfies

$$\Pr[S \geq (1+\epsilon)\mu] \leq e^{\frac{-\epsilon^2 \mu}{3}}.$$

and for $0 < \epsilon < 1$

$$\Pr[S \leq (1-\epsilon)\mu] \leq e^{\frac{-\epsilon^2 \mu}{2}}.$$

Quantitative versions of the CLT

Theorem (Bernstein Inequality)

Let $X_1, X_2, \ldots, X_k$ be independent random variables with each $X_i \in [-1, 1]$. Let $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = \mathrm{Var}[X_i]$. Let $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Then, for $k \leq \frac{1}{2}\sigma$, $S = \sum_i X_i$ satisfies

$$\Pr[|S - \mu| > k \cdot \sigma] \leq 2 \exp\left(-\frac{k^2}{4}\right).$$

Quantitative versions of the CLT

Theorem (Hoeffding Inequality)

Let $X_1, X_2, \ldots, X_k$ be independent random variables with each $X_i \in [a_i, b_i]$. Let $\mu_i = \mathbb{E}[X_i]$ and $\mu = \sum_i \mu_i$. Then, for any $\alpha > 0$, $S = \sum_i X_i$ satisfies:

$$\Pr[|S - \mu| > \alpha] \leq 2 \exp\left(-\frac{\alpha^2}{\sum_{i=1}^{k}(b_i - a_i)^2}\right).$$

Chernoff Bound application

Sample Application: Flip biased coin k times: i.e. the coin is heads with probability b. As long as $k \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$,

$b \in [0, 1]$ $\quad \Pr[|\#\text{ heads} - b \cdot k| \geq \epsilon k] \leq \delta$

Setup: Let $X_i = \mathbb{1}[i^{\text{th}}$ flip is heads$]$. Want bound probability that $\sum_{i=1}^{k} X_i$ deviates from it's expectation.

Corollary of Chernoff bound: Let $S = \sum_{i=1}^{k} X_i$ and $\mu = \mathbb{E}[S]$. For $0 < \Delta < 1$,

$$\Pr[|S - \mu| \geq \Delta\mu] \leq 2e^{-\Delta^2\mu/3}$$

Chernoff Bound application

Sample Application: Flip biased coin k times. i.e. the coin is heads with probability b. As long as $k \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$,

$$\Pr[|\#\text{ heads} - b \cdot k| \geq \epsilon k] \leq \delta$$

Pay very little for higher probability – if you increase the number of coin flips by 2x, δ goes from $1/10 \to 1/100 \to 1/10000$

Application: Median Trick

A even better trick than taking the mean:

Lemma

Let $Z_1, Z_2, \ldots, Z_t$ be random "estimates" of some unknown value $R \in \mathbb{R}$, such that the Z_i's are i.i.d. and such that $\Pr[|Z_i - R| < \epsilon] \geq \frac{2}{3}$ for each $i \in [t]$, for any $\epsilon > 0$.

Then, for any $\delta \in (0,1)$, setting $t = O(\log \frac{1}{\delta})$, we have:

$$\Pr\left[\left|\text{MEDIAN}_{i \in [t]} Z_i - R\right| \leq \epsilon\right] \geq 1 - \delta$$

Proof: Define indicator variables $X_i = 1$ if $|Z_i - R| < \epsilon$. Then $X_1, \ldots, X_t$ are independent "coin flips" with mean $b > 2/3$. By Chernoff, $\sum_i X_i > (1 - \frac{1}{10})\frac{2}{3} > \frac{1}{2}$ with probability $1 - \delta$!

Application: Median Trick

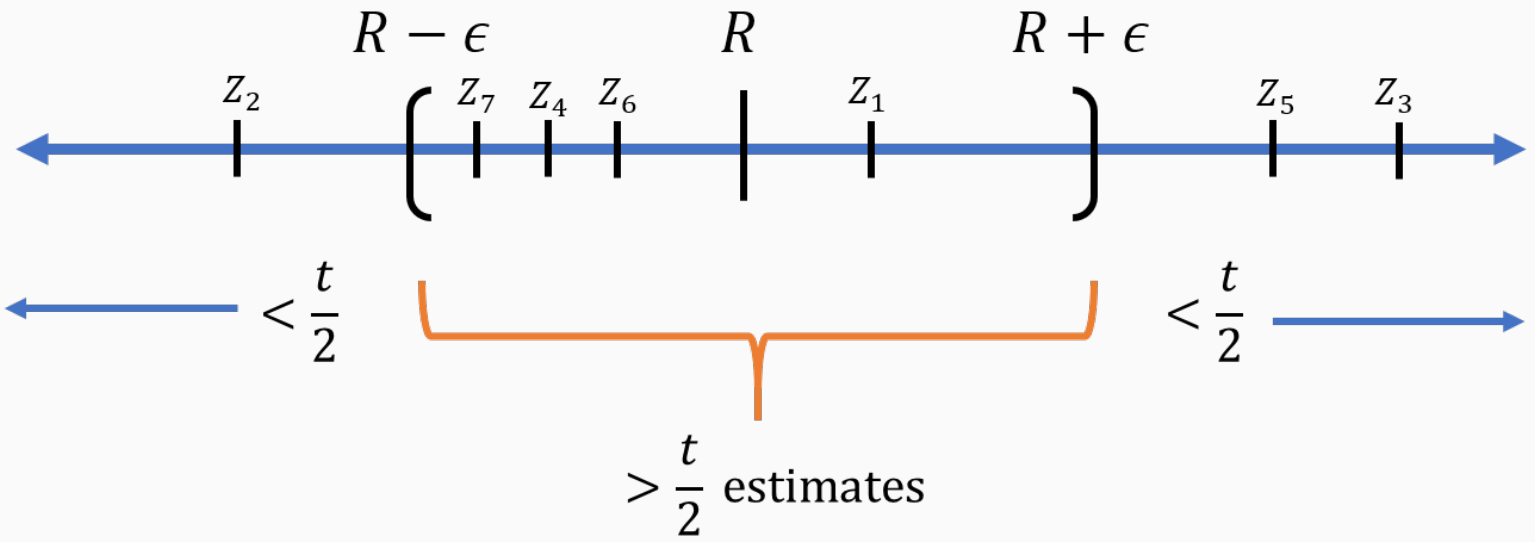
Proof: Define indicator variables $X_i = 1$ if $|Z_i - R| < \epsilon$. Then $X_1, \ldots, X_t$ are independent "coin flips" with mean $b > 2/3$. By Chernoff, $\sum_i X_i > \frac{1}{2}$ with probability $1 - \delta$!

The Neat Observation: Given estimates $Z_1, \ldots, Z_t$ of R, if $> 1/2$ of the estimates satisfy $|Z_i - R| < \epsilon$, then we have $\left|\text{MEDIAN}_{i \in [t]} Z_i - R\right| \leq \epsilon$

Application: Median Trick

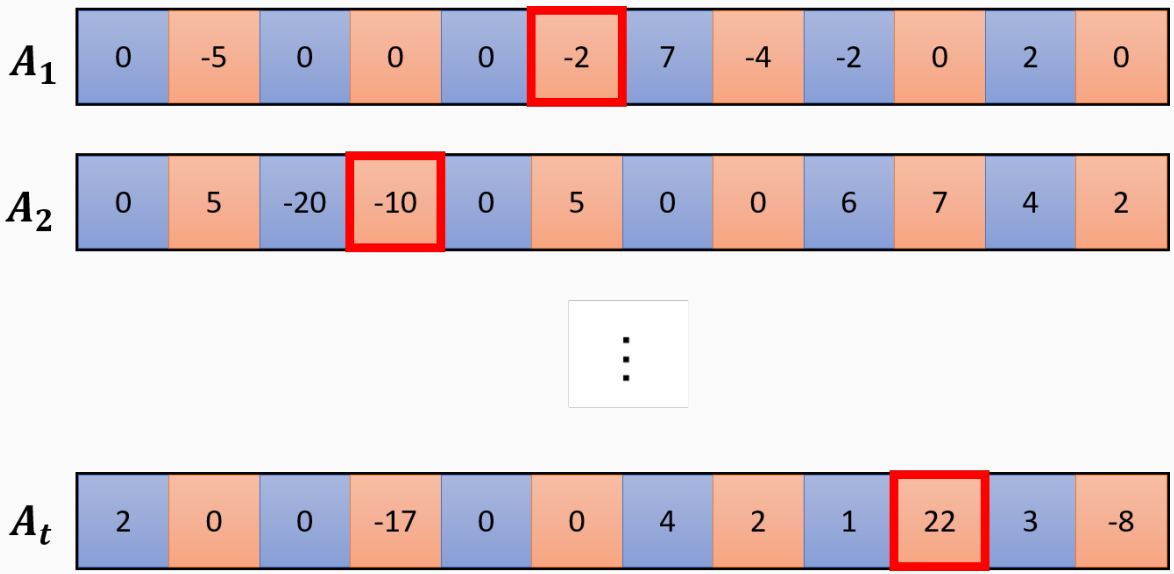
Proof: Define indicator variables $X_i = 1$ if $|Z_i - R| < \epsilon$. Then $X_1, \ldots, X_t$ are independent "coin flips" with mean $b > 2/3$. By Chernoff, $\sum_i X_i > \frac{1}{2}$ with probability $1 - \delta$!

The Neat Observation: Given estimates $Z_1, \ldots, Z_t$ of R, if $> 1/2$ of the estimates satisfy $|Z_i - R| < \epsilon$, then we have $\left|\text{MEDIAN}_{i \in [t]} Z_i - R\right| \leq \epsilon$



Back to Count-Sketch

A Better Estimator: $\tilde{f}_i = \text{MEDIAN}_{j\in[t]}\sigma_j(i)A_j[h_j(i)]$



Back to Count-Sketch

A Better Estimator: $\tilde{f}_i = \text{MEDIAN}_{j\in[t]}\sigma_j(i)A_j[h_j(i)]$

$$Z_j = \sigma_j(i)A_j[h_j(i)] = f_i + \underbrace{\sum_{k\neq i} \mathbb{1}[\mathbf{h}(k) = \mathbf{h}(i)] \cdot f_k}_{\text{error in frequency estimate}}$$

We showed $\mathbb{E}[\text{error}] = 0$ and $\text{Var}[\text{error}] < \frac{1}{B}\|f\|_2^2$, so by Chebyshev's, setting $B = \Theta(1/\epsilon^2)$ we have

$$\Pr[|Z_i - f_i| < \epsilon\|f\|_2] \geq \frac{2}{3}$$

Exactly the same set-up as the earlier Lemma! Each estimate Z_i is correct independently with probability at least $2/3$.

Back to Count-Sketch

A Better Estimator: $\tilde{f}_i = \text{MEDIAN}_{j\in[t]}\sigma_j(i)A_j[h_j(i)]$

Theorem

Fix any $\epsilon, \delta \in (0,1)$. Then for any $i \in [n]$, the **Count-Sketch** *algorithm, using the estimate above, satisfies*

$$\Pr\left[|\tilde{f}_i - f_i| > \epsilon\|f\|_2\right] \leq \delta$$

Using space $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$.

This was our goal probability!

Back to Count-Sketch

Setting $\delta = 1/n^2$ in the earlier example, and applying the union bound, we have:

Theorem

Fix any $\epsilon \in (0, 1)$. *Then* **simultaneously for every** $i \in [n]$, *the* **Count-Sketch** *algorithm satisfies*

$$\Pr\left[|\tilde{f}_i - f_i| > \epsilon \|f\|_2\right] \leq \frac{1}{n}$$

Using space $O(\frac{1}{\epsilon^2} \log n)$.

Since each estimate is good, we can scan through all the estimates $\{\tilde{f}_i\}_{i \in [n]}$, and return the set $S = \{i \ : \ \tilde{f}_i > 3\epsilon \|f\|_1\}$.

This solves the 4ϵ-Heavy Hitters problem in $O(\frac{1}{\epsilon^2} \log n)$ space!

Another Application Of The Union Bound: Balls-in-Bins

Load Balancing

Load balancing problem:

Suppose Google answers map search queries using servers $A_1, \ldots, A_q$. Given a query like "new york to rhode island", common practice is to choose a random hash function $h \to \{1 \ldots, q\}$ and to route this query to server:

$$A_{h(\text{"new york to rhode island'})}$$

Why use a hash function instead of just distributing requests randomly?

Goal: Ensure that requests are distributed evenly, so no one server gets loaded with too many requests. We want to avoid downtime and slow responses to clients.

Load Balancing

Suppose we have n servers and m requests, $x_1, \ldots, x_m$. Let s_i be the number of requests sent to server $i \in \{1, \ldots, n\}$:

$$s_i = \sum_{j=1}^{m} \mathbb{1}[h(x_j) = i].$$

Formally, our goal is to understand the value of maximum load on any server, which can be written as the random variable:

$$S = \max_{i \in \{1,\ldots,n\}} s_i.$$

Load Balancing

A good first step in any analysis of random variables is to first think about expectations. If we have n servers and m requests, for any $i \in \{1, \ldots, n\}$:

$$\mathbb{E}[s_i] = \sum_{j=1}^{m} \mathbb{E}\left[\mathbb{1}[h(x_j) = i]\right] = \frac{m}{n}.$$

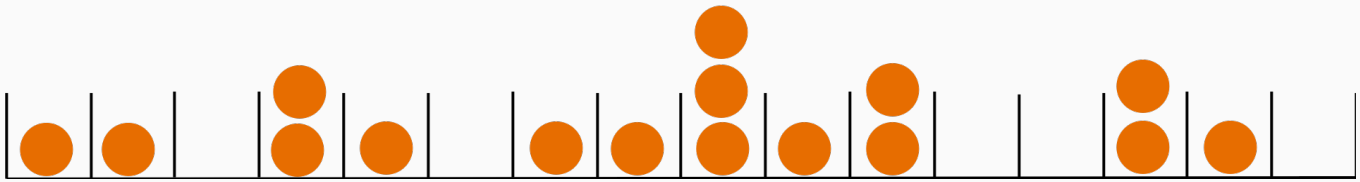
But it's very unclear what the expectation of $S = \max_{i \in \{1,\ldots,n\}} s_i$ is... in particular, $\mathbb{E}[S] \neq \max_{i \in \{1,\ldots,n\}} \mathbb{E}[s_i]$.

Exercise: Convince yourself that for two random variables A and B, $\mathbb{E}[\max(A, B)] \neq \max(\mathbb{E}[A], \mathbb{E}[B])$ even if those random variable are independent.

Balls-into-bins

Number of servers: To reduce notation and keep the math simple, let's assume that $m = n$. I.e., we have exactly the same number of servers and requests.

Hash function: Continue to assume a fully (uniformly) random hash function h.



Often called the "balls-into-bins" model.

$\mathbb{E}[s_i]$ = expected number of balls per bin $= \frac{m}{n} = 1$. We would like to prove a bound of the form:

$$\Pr[\max_i s_i \geq C] \leq \frac{1}{10}.$$

for as tight a value of C. I.e., something much better than $C = n$.

Application of Union Bound

We want to prove that:

$$\Pr[\max_i s_i \geq C] = \Pr[(s_1 \geq C) \cup (s_2 \geq C) \cup \ldots \cup (s_n \geq C)] \leq \frac{1}{10}.$$

To do so, it suffices to prove that for all i:

$$\Pr[s_i \geq C] \leq \frac{1}{10n}.$$

Why? Because then by the union bound,

$$\Pr[\max_i s_i \geq C] \leq \sum_{i=1}^{n} \Pr[s_i \geq C] \quad \text{(Union bound)}$$

$$\leq \sum_{i=1}^{n} \frac{1}{10n} = \frac{1}{10}. \quad \square$$

High probability bounds

Prove that for some C,

$$\Pr[s_i \geq C] \leq \frac{1}{10n}.$$

This should look hard! We need to prove that $s_i < C$ (i.e. the i^{th} bin has a small number of balls) with very high probability (specifically $1 - \frac{1}{10n}$).

Markov's inequality is too weak of a bound for this.

n = number of balls and number of bins. s_i is number of balls in bin i.
C = upper bound on maximum number of balls in any bin.

Better Concentration:

Exercise: Show that $\mathrm{Var}(s_i) \leq 1$.

- Using Chebyshev's Inequaltity, we obtain $\Pr[s_i > 10\sqrt{n}] \leq \frac{1}{100n}$.
- Union bound gives $\Pr[\max_i s_i \geq 10\sqrt{n}] \leq \frac{1}{100}$.
- Chebyshev's gives us a max load of $O(\sqrt{n})$, can we do better?

Better Concentration:

Exercise: Show that $\mathrm{Var}(s_i) \leq 1$.

- Using Chebyshev's Inequaltity, we obtain $\Pr[s_i > 10\sqrt{n}] \leq \frac{1}{100n}$.
- Union bound gives $\Pr[\max_i s_i \geq 10\sqrt{n}] \leq \frac{1}{100}$.
- Chebyshev's gives us a max load of $O(\sqrt{n})$, can we do better?

Exercise: $s_i = \sum_j s_{i,j}$, where $s_{i,j} := j$-th ball lands in i-th bin
Then $s_{i,j}$'s are i.i.d. indicator random variables. Use Chernoff bound to show that $\Pr[s_i > 100 \log(n)] \leq \frac{1}{n^2}$.

- Chernoff gives a max load of $O(\log n)$!
- Can actually do even better, and get a max load of $O(\log n / \log \log n)$!

A simple proof

Will show $\Pr[s_i > 6\frac{\log n}{\log\log n}] < 1/n^2$.

Proof: Set $C = 6\frac{\log n}{\log\log n}$, then $\Pr[s_i > C] = \sum_{k \geq C} \Pr[s_i = k]$, so for any fixed k:

$$\Pr[s_i = k] = \sum_{k \geq C} \binom{n}{k} \cdot \left(\frac{1}{n}\right)^k \leq \sum_{k \geq C} \left(\frac{en}{k}\right)^k \left(\frac{1}{n}\right)^k$$

$$\leq \left(\frac{e}{k}\right)^k \leq \left(\frac{\log\log n}{\log n}\right)^{\frac{6\log n}{\log\log n}} \leq \left(\frac{1}{\sqrt{\log n}}\right)^{\frac{6\log n}{\log\log n}}$$

$$\leq 2^{-\frac{1}{2}\log(\log n)\frac{6\log n}{\log\log n}} \leq \frac{1}{n^3}$$

So $\Pr[s_i > C] = \sum_{k > C} \Pr[s_i = k] \leq \frac{1}{n^2}$. This is actually tight!

Takeaways

Techniques used that will appear again:

- Use exponential concentration inequalities (or direct calculations) to get tight bounds on probability of an individual random variable.
- Then apply the union bound to control the maximum of a collection of such variables.

Next Class: The celebrated Johnson-Lindenstrauss Lemma and High-Dimensional Geometry.